JOURNAL OF APPLIED BEHAVIOR ANALYSIS 2011, **44,** 937–941 NUMBER 4 (WINTER 2011)

INTERTEACHING: THE IMPACT OF LECTURES ON STUDENT PERFORMANCE

BRYAN K. SAVILLE, TROY COX, SEAN O'BRIEN, AND ARIANA VANDERVELDT

JAMES MADISON UNIVERSITY

Several studies suggest that interteaching improves student learning more than traditional lectures, but few have examined which components of interteaching contribute to its efficacy. We examined whether the lecture component of interteaching affected students' exam grades and cumulative point totals in a research methods course. Although students who received lectures had consistently higher exam scores than students who did not, the differences were statistically significant on only 2 of 5 exams. Students who received lectures, however, earned significantly more points during the semester.

Key words: college teaching, exam scores, interteaching, lecture

Interteaching is an approach to classroom instruction that has its roots in behavior analysis (Boyce & Hineline, 2002). A typical interteaching session proceeds as follows. Before each class, the instructor distributes a preparation (prep) guide that contains questions designed to lead students through a reading assignment. Students answer the questions before class and prepare to discuss their answers with another student. Each class begins with a lecture, in which the instructor spends one third of the period reviewing material from the last class that students found difficult. After the lecture, students form pairs and spend the remaining time discussing their completed prep guides. The instructor moves among the pairs, answering questions during the discussions. When students finish their discussions, they submit a record sheet on which they list any questions they would like reviewed. The instructor uses this information to prepare a clarifying lecture that begins the next class period. Students receive a small number of points for participating in the discussions and "quality" points if both they and their discussion partners do well on the exams.

Since Boyce and Hineline's (2002) initial description of interteaching, several studies have found it to be more effective than traditional lectures (Saville, Lambert, & Robertson, 2011). Fewer studies, however, have examined which components contribute to its efficacy. In one study, Saville and Zinn (2009) found that quality points did not affect exam scores. In another study, Cannella-Malone, Axe, and Parker (2009) found that exam scores were similar when college students developed and answered their own prep-guide questions or completed instructor-prepared prep guides.

Another component that is ripe for analysis is the lecture component of interteaching. It seems reasonable that the lectures might affect student performance because they target material that students report as being difficult. The purpose of this study was to examine the impact of lectures on two measures of performance: students' exam scores and students' final grades, as determined by their cumulative point totals on the exams.

METHOD

Participants and Setting

Participants were 46 undergraduate students (median age = 20 years) from James Madison

Ariana Vanderveldt is now in the Department of Psychology at Washington University in St. Louis.

We thank Tracy Zinn for her comments on an earlier version of this manuscript.

Correspondence concerning this article should be addressed to Bryan K. Saville, Department of Psychology, James Madison University, Harrisonburg, Virginia (e-mail: savillbk@jmu.edu).

doi: 10.1901/jaba.2011.44-937

University, who were enrolled in one of three sections of a research methods course taught by the first author. There were 15 students (two men, 13 women) in Section 1; 15 students (five men, 10 women) in Section 2; and 16 students (three men, 13 women) in Section 3. The sections met for 75 min on Tuesdays and Thursdays (although students in Section 3 did not always stay for the full 75 min; see below).

Materials and Procedure

Because we could not randomly assign participants to the sections, we collected the following self-reported demographic data during the first week of class: (a) gender, (b) age, (c) cumulative grade-point average (GPA), (d) number of psychology courses already completed, (e) number of credit hours taken during the semester, (f) grade in a prerequisite statistics course, and (g) employment status. These data helped us to determine whether the sections were similar before our manipulation.

Students in each section completed an instructor-prepared prep guide before each class. Each prep guide contained eight to 12 items, and each item usually included two or more questions. The items were in short-answer and essay format and typically required students to explain ideas (e.g., "Variables that are valid are likely reliable, but variables that are reliable are not necessarily valid.") or apply concepts to research problems (e.g., "In the following scenario, what is one variable that might confound the results?").

During each class, students formed pairs and had 45 to 55 min to discuss the prep guides. If students took fewer than 45 min to finish, the instructor recommended that they continue to discuss and review their answers. Students were free to choose their own partners but were instructed not to work with the same person more than three times during the semester. During discussions, the instructor and a teaching assistant (TA) moved among the pairs, answering questions and guiding the discussions if students were confused. In general, though, the discussions were driven by students' comments to one another and not by periodic instructor questioning to the entire class. After students finished their discussions, they submitted a record sheet on which they listed their partner's name, how well the discussion went, and which questions they would like reviewed (the record sheets for Section 3 did not contain the last item; see below). For each discussion they completed, students earned points that, across the semester, totaled 10% of their course grades.

To examine the impact of lectures, we exposed each section to a different lecture condition. Students in Section 1 (delayed lecture) received their lectures at the start of the next class period, either 2 or 5 days later, depending on whether the discussions took place on Tuesday or Thursday. Students in Section 2 (immediate lecture) received lectures approximately 5 min after they finished their discussions and submitted the record sheets. In each of the lecture conditions, the instructor lectured for 20 to 30 min, reviewing the three or four prep-guide items listed most frequently on the record sheets and answering any additional questions. Students in Section 3 (control) did not receive lectures; rather, after completing their discussions and submitting their record sheets, they were free to leave.

Students in each section took the same 30-point exam after each unit of information (usually three or four prep guides). Each exam consisted of three 5-point essay questions and other objective questions that required students to solve problems (e.g., "What is the level of interobserver reliability in this scenario?"), apply information (e.g., "Identify the threat to internal validity in this scenario."), and show higher level comprehension of the material covered on the prep guides (e.g., "Design a study to examine the effects of visualization on free-throw shooting in college students."). Students took five exams during the semester.

Although students knew whether they would be hearing lectures during the semester (this information was in the course syllabus and

discussed the first day of class), they were not initially told the purpose of the study. On the last day of class, we informed students of the purpose of the study and then asked them to read and sign consent forms allowing us to use their data. Except for one student in the control condition, who dropped the class after Exam 3, every student provided consent.

Response Measurement and Interobserver Agreement

For each of the exams, a TA first graded the entire set. A second TA then graded a randomly chosen subset of 15 exams (about 33% of the total set). The TAs were blind to the group assignment while grading. To ensure independence in grading, the TAs placed their scoring for each of the 15 exams on separate sheets of paper. We calculated interobserver agreement by taking the number of items on which the TAs agreed (i.e., assigned the same number of points for an answer), dividing it by the total number of items on the exam, and converting the ratio to a percentage. Agreement scores across the five exams ranged from 73% to 97%, with a mean score of 88%. Disagreements usually occurred on the essay questions on which the TAs' grading varied by a half point. When there were disagreements, the TAs discussed their grading and came to an agreement on the final score.

RESULTS AND DISCUSSION

Two chi-square tests (gender, employment), one Kruskal-Wallis test (statistics grade), and four independent-samples t tests (age, GPA, number of psychology courses, and credit hours) found no significant differences among sections on any of the demographic measures (all $ps > .10$). As noted above, one student from the control condition dropped the class after Exam 3. We thus removed her data, leaving 15 students in each condition for the following analyses.

We first examined whether there were significant differences among conditions on each of the five exams (Figure 1). An independent-samples ANOVA found significant differences on Exam 1, $F(2, 42) = 9.13$, $p = .001$, $\eta_p^2 = 0.3$, and Exam 3, $F(2, 42) = 7.63$, $p = .001$, $\eta_p^2 = 0.27$. On Exam 1, a Sidak post-hoc test showed that students in the delayed-lecture condition ($M = 85\%$) had significantly higher exam scores ($p < .001$, $d = 1.61$) than students in the control condition ($M = 71\%$), but were not significantly different ($p = .10$) from students in the immediate-lecture condition ($M = 78\%$). There was also no significant difference ($p = .13$) between students in the immediate-lecture and control conditions. On Exam 3, a Sidak post-hoc test showed that students in the delayed-lecture ($M = 88\%$) and immediate-lecture ($M = 89\%$) conditions had significantly higher exam scores ($ps = .006$ and .004, respectively; $ds = 1.31$ and 1.20, respectively) than students in the control condition ($M = 78\%$), but were not different from one another ($p = .99$). Finally, on Exams 2, 4, and 5, although students in the lecture conditions had consistently higher scores than students in the control condition, none of the differences were significant ($ps > .25$).

The lack of consistent significant differences makes it difficult to determine whether the lectures affected students' exam scores and, thus, whether lectures are an important component of interteaching. Given, however, that students in the lecture conditions had consistently higher mean scores than students in the control condition, it seems possible that small nonsignificant differences on the individual exams might have contributed to larger significant differences that emerged across the semester.

To examine whether there were cumulative differences in student performance, we calculated the total number of points (of a possible 150) that students earned across the five exams. An independent-samples ANOVA found significant differences among the groups, $F(2, 42)$

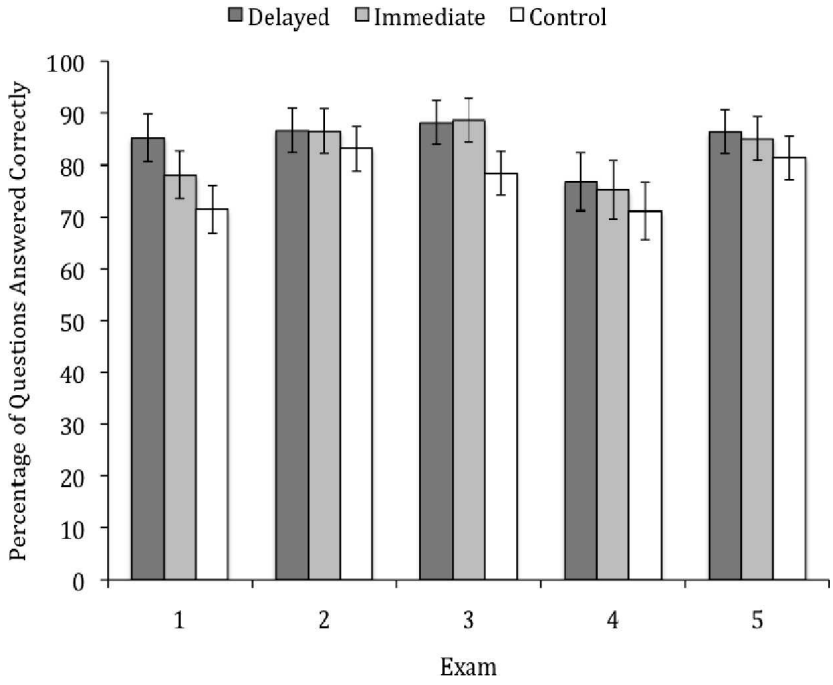

Figure 1. Mean percentage of questions answered correctly on each exam by students in the delayed-lecture (dark gray), immediate-lecture (light gray), and control (white) conditions. Error bars represent 95% confidence intervals.

= 5.38, $p = .008$, $\eta_p^2 = 0.2$. A Sidak post-hoc test showed that students in the delayed-lecture ($M = 127$ points) and immediate-lecture ($M = 124$ points) conditions earned significantly more points (ps = .009 and .05, respectively; ds = 1.22 and 0.84, respectively) than students in the control condition ($M = 115$ points) but were not significantly different from one another ($p = .81$).

In sum, the reliably higher exam performance of students in the lecture conditions appears to have contributed to cumulative differences that emerged across the semester. When converted into typical college grades, these differences were equivalent to students in the lecture conditions earning a B (83% to 85%), on average, and students in the control condition earning a C (77%), a difference that might be important to some, if not many, students and educators. From our second analysis, then, one might conclude that the lectures are important to interteaching because they produce practical differences (Kirk, 1996) that affect students' cumulative performance. Moreover, whether the lectures were delayed or immediate did not seem to matter, which may be important both to students and instructors. For example, some students might prefer immediate lectures because they like to leave each class period knowing the answers to questions they had; others might prefer delayed lectures because they have a few days to think about the material before receiving clarification. Similarly, experienced instructors who are familiar with the material may prefer to "wing it" and lecture immediately after the discussions, whereas less experienced instructors may prefer to have more time to review the material before presenting it. Although previous research (Saville, Zinn, Neef, Van Norman, & Ferreri, 2006) has examined students' preferences for interteaching, no researchers have studied instructors' preferences. Thus, it might be useful to examine both students' and instructors' preferences for immediate and

delayed lectures as well as instructors' preferences for interteaching in general.

This study has at least three limitations. First, because we had 15 students per condition, the lack of significant differences on some of the exams may have been a function of low statistical power. Second, students in the lecture conditions may have performed better on the exams simply because they had additional exposure to the material (via the lectures). Finally, although our results provide some evidence that lectures are important to interteaching, we could not determine whether clarifying lectures (i.e., those based on student comments) were responsible for the differences. It is certainly possible that any type of lecture, regardless of whether it is based on student requests, may improve performance in interteaching courses. Moreover, experienced instructors may target material that, historically, has been difficult for students, regardless of whether current students have similar difficulties. In this way, experienced instructors' nonclarifying lectures may function in much the same way as the clarifying lectures delivered by more inexperienced instructors. Future research should continue to examine whether lectures in general, or clarifying lectures more specifically, contribute to interteaching outcomes.

REFERENCES

Boyce, T. E., & Hineline, P. N. (2002). Interteaching: A strategy for enhancing the user-friendliness of behavioral arrangements in the college classroom. *The Behavior Analyst*, *25*, 215–226.

Cannella-Malone, H. I., Axe, J. B., & Parker, E. D. (2009). Interteach preparation: A comparison of the effects of answering versus generating study guide questions on quiz scores. *Journal of the Scholarship of Teaching and Learning*, *9*, 22–35.

Kirk, R. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746–75.

Saville, B. K., Lambert, T., & Robertson, S. (2011). Interteaching: Bringing behavioral education into the 21st century. *The Psychological Record*, *61*, 153–166.

Saville, B. K., & Zinn, T. E. (2009). Interteaching: The effects of quality points on exam scores. *Journal of Applied Behavior Analysis*, *42*, 369–374.

Saville, B. K., Zinn, T. E., Neef, N. A., Van Norman, R., & Ferreri, S. J. (2006). A comparison of interteaching and lecture in the college classroom. *Journal of Applied Behavior Analysis*, *39*, 49–61.

Received October 11, 2010
Final acceptance January 14, 2011
Action Editor, John Borrero